PREDICT: Multi-Agent-based Debate Simulation for Generalized Hate Speech Detection

Trigger Warning: This paper contains discussions of hate speech that may be distressing or triggering for some readers.

Someen Park[1], Jaehoon Kim[1], Seungwan Jin[2], Sohyun Park[1], Kyungsik Han[1,2,*]
[1] Department of Artificial Intelligence, Hanyang University, Seoul, Republic of Korea
[2] Department of Data Science, Hanyang University, Seoul, Republic of Korea
{someeeen,jaehoonkimm,seungwanjin,sohyunpark,kyungsikhan}@hanyang.ac.kr

Abstract

While a few public benchmarks have been proposed for training hate speech detection models, the differences in labeling criteria between these benchmarks pose challenges for generalized learning, limiting the applicability of the models. Previous research has presented methods to generalize models through data integration or augmentation, but overcoming the differences in labeling criteria between datasets remains a limitation. To address these challenges, we propose **PREDICT**, a novel framework that uses the notion of multi-agent for hate speech detection. PREDICT consists of two phases: (1) **PRE** (**P**erspective-based **RE**asoning): Multiple agents are created based on the induced labeling criteria of given datasets, and each agent generates stances and reasons; (2) **DICT** (**D**ebate using **I**n**C**ongruen**T** references): Agents representing hate and non-hate stances conduct the debate, and a judge agent classifies hate or non-hate and provides a balanced reason. Experiments on five representative public benchmarks show that PREDICT achieves superior cross-evaluation performance compared to methods that focus on specific labeling criteria or majority voting methods. Furthermore, we validate that PREDICT effectively mediates differences between agents' opinions and appropriately incorporates minority opinions to reach a consensus. Our code is available at https://github.com/Hanyang-HCC-Lab/PREDICT

1 Introduction

The rise of hate speech on the Internet has become a significant social issue, prompting extensive research on hate speech detection (Moy et al., 2021; Jahan and Oussalah, 2023; Zhou et al., 2021). One of the main difficulties encountered in hate speech detection is generalization (Yin and Zubiaga, 2021). This refers to the situation where an effective model trained on a particular dataset may perform poorly when the model is applied to a different dataset (Cai et al., 2022). This is mainly due to differences in various labeling criteria (Ramalingam et al., 2022), including the purpose (identifying hate speech in social and historical contexts, classifying sentiments to classify hate speech, or focusing on the targets of hate speech), the labeling method (the number of annotators, the labeling process, and guidelines), the granularity of the labels (multi-class or binary), the size of the dataset, and the time and method of data collection (Sachdeva et al., 2022; Khurana et al., 2022). Thus, an approach that does not overly rely on specific labeling criteria is needed to improve the generalization of hate speech detection.

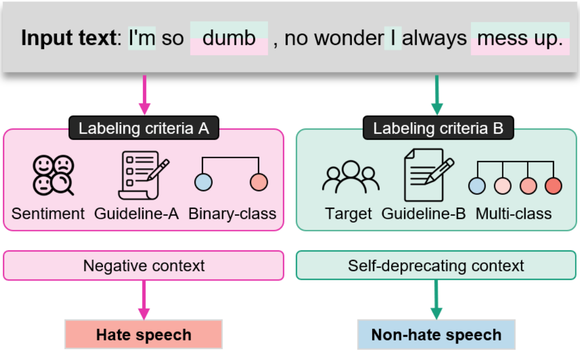

Figure 1: Our research is motivated by the classification of the same text under different labeling criteria. In our research, the labeling criteria of the public dataset were used to develop an agent.

Previous studies have attempted to address the issue of generalization in hate speech detection through various approaches, including data integration, augmentation, and explanation generation. The integration of datasets covering diverse topics, such as gender and race (Bourgeade et al., 2023), allows the model to learn more extensive hate speech patterns. However, differences in the labeling of the datasets can cause sentences with similar words

[*] Corresponding author

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 20963–20987
November 12-16, 2024 ©2024 Association for Computational Linguistics

or expressions to be labeled differently, which can confuse the model regarding consistency. The recent data augmentation through GPT-2 (Wullach et al., 2021) allows the model to learn various forms of hate speech. However, this approach has a limitation in that it may result in the generation of repetitive patterns that exist in the original datasets. The use of GPT-generated explanations in training (Yang et al., 2023) also depends on specific criteria for labeling data, resulting in underperformance on datasets based on different labeling criteria. While these approaches of previous research have partially improved the performance of hate speech detection, there remain limitations in effectively incorporating differences in labeling criteria into the model's training or inference process.

Our research employs a pluralistic approach (Waseem et al., 2018) to build consensus based on respect and inclusion of diverse perspectives and to address the overfitting caused by different labeling criteria across datasets. Recent social science research emphasizes the value of pluralism in addressing hate speech debates (Tontodimamma et al., 2021), advocating the inclusion of diverse values and perspectives and highlighting the need for rational integration of these views to achieve social consensus (Dudley-Marling and Burns, 2014; Feldman, 2021).

In this paper, we present a PREDICT framework that uses a Large Language Model (LLM) to construct a debate environment among different agents, where each agent has its independent perspectives based on the labeling criteria of the hate speech dataset, and simulates pluralistic decision-making in hate speech detection. PREDICT consists of two phases: the "Perspective-based REasoning (PRE)" phase and the "Debate using InCongruenT references (DICT)" phase. The former is designed to form a perspective based on the assigned labeling criteria and similar contexts. The latter, in turn, simulates a debate between two debaters, with the judge providing the final label and justification. The DICT phase is structured to reach a final decision through two rounds of debate motivated by Liang et al. (2023) and Xiong et al. (2023).

We conducted experiments on five public benchmarks (five different agents) in the hate speech domain to quantitatively evaluate the performance of PREDICT in the context of generalized hate speech detection (Section 4). Our results show the significant effectiveness of PREDICT in accurately detecting hate speech in all five datasets and in deriving a strong justification for the decision.

In summary, our study highlights the importance of consensus in hate speech research and demonstrates the value of multiple perspectives as a way to improve the accuracy of hate speech detection. Our contributions are as follows:

- **Respect Diverse Perspective**: By defining differences in the labeling criteria of diverse datasets as "independent perspectives," PREDICT respects diverse perspectives on hate speech and stores them as a reference for debate.
- **Consensus through Debate**: PREDICT presents a reasoning-based debate simulation for hate speech detection that facilitates consensus among multi-agent, each of whom is assigned an independent perspective.
- **Generalization**: We validate the generalization of the proposed PREDICT and its superior hate speech detection performance.

2 Related Work

2.1 Hate Speech Detection Generalization

The challenge of generalized hate speech detection has been addressed by various approaches (Rizos et al., 2019; Ludwig et al., 2022; Pendzel et al., 2024; Nirmal et al., 2024; Fortuna et al., 2020; Jin et al., 2023). We categorize and explain three primary approaches such as data augmentation, data integration, and explanation generation via LLMs.

Wullach et al. (2021) proposed a method for generating a substantial amount of synthetic hateful text, using the GPT-2 on a specific hate speech dataset. This approach enhanced hate speech detection by increasing the size of the dataset and addressing imbalances between the hate and non-hate data. However, it was limited by the fact that the generated data still reflected the biases present in the original data.

Bourgeade et al. (2023) integrated a dataset of hate speech covering a wide range of topics, including gender and race. The model was trained on this dataset, learning patterns in different forms of hate speech. This allowed the model to understand the relationships between different topics and to reduce its bias towards certain topics. However, differences in the annotation guidelines across datasets led to significant inconsistent labeling of the same topic, limiting the model's ability to generalize.

Yang et al. (2023) used LLM to generate free text rationales of hate speech through chain-of-thought prompts. The generated rationales enabled the model to better understand the nuances and context of the text, thereby improving its accuracy in detecting hate speech. However, the lack of common features across various datasets, such as linguistic patterns (specific word usage, sentence construction) and structural elements (sentence length, dialog format), made generalization challenging.

Hong and Gauch (2023) proposed a multi-task learning framework that simultaneously trained hate speech detection (primary task) and sentiment analysis (secondary task). The framework used a shared parameter encoder to facilitate knowledge sharing between the two tasks and investigated the use of incorporating additional sentiment labels to improve the generalization of the hate speech detection model. The limited quality and range of sentiment data made this approach effective only on certain datasets.

Despite various technical approaches to generalized hate speech detection, the insufficient consideration of differences in labeling criteria across datasets remains a significant challenge. This indicates that further efforts are needed to recognize and overcome the differences between datasets. In this study, we adopt a pluralistic approach to effectively address the challenges posed by differences in labeling criteria across datasets and explore ways to improve the generalization performance of hate speech detection.

2.2 Agent Debate

In significant advances in LLMs, research has been conducted to improve their performance in certain downstream tasks, such as arithmetic problems and translation, by having multiple agents simulate human behavior (Liang et al., 2023; Wu et al., 2023; Subramaniam et al., 2023).

Chan et al. (2023) proposed CHATEVAL, a system for evaluating the quality of LLM-generated answers to questions on various domains and topics. CHATEVAL employed a multi-agent approach to evaluate the answers, thereby increasing the accuracy and reliability of assessments for given questions. Du et al. (2023) proposed a method for multiple agents to independently analyze and derive solutions to given tasks, such as arithmetic problems. Agents critically review and debate each other's proposals, presenting counterarguments that improve the factuality and reasoning capabilities of the language models. Liang et al. (2023) offered a framework, MAD, that sets up agents as proponents and opponents, enabling the agents to argue and debate the conclusions generated by LLMs from specific tasks, such as translation and arithmetic problems. As demonstrated by MAD, the problem of self-reflection-based prompt engineering, as defined as Degeneration of Thoughts (DoT), can be mitigated by providing external feedback to each other.

Previous studies have demonstrated the effectiveness of LLM-based multi-agent interaction and debate frameworks in various downstream tasks. However, the generation of text by agents for interaction still depends on the underlying internal knowledge of the LLMs (Gallegos et al., 2024; Cai et al., 2024; Kumar et al., 2024). This simple approach is limited in providing an objective and consistent evaluation due to the inherent biases and uncertainties of the agents. In contrast, in the domain of hate speech detection, it is essential to perform consistent and unbiased simulations of the debate.

In this paper, we manually refine the labeling criteria of each public benchmark based on prior research and assign them to each agent's independent perspective. This approach allows us to propose PREDICT, a framework for generating and debating arguments and reasons based on each agent's perspective. To the best of our knowledge, our work is the first to present a multi-agent-based simulation in the domain of hate speech.

2.3 Theoretical Background of PREDICT

Various theoretical and empirical studies provide a foundation for how multi-agent systems can improve decision-making in hate speech detection. Mannes et al. (2014) showed that group averages are more accurate than individual estimates. This may explain how the multi-agent collective judgments of the PREDICT framework can overcome the limitations of single agent judgments and lead to more accurate hate speech detection. Bose et al. (2017) emphasized that groups can make effective decisions without centralized leadership. This theory is central to the PREDICT framework, where agents reason independently and reach consensus. This approach avoids relying on a single agent who specializes in a specific dataset, which may otherwise resemble centralized leadership.

Davani et al. (2023) showed how the biases of human annotators influence AI systems. These bi-

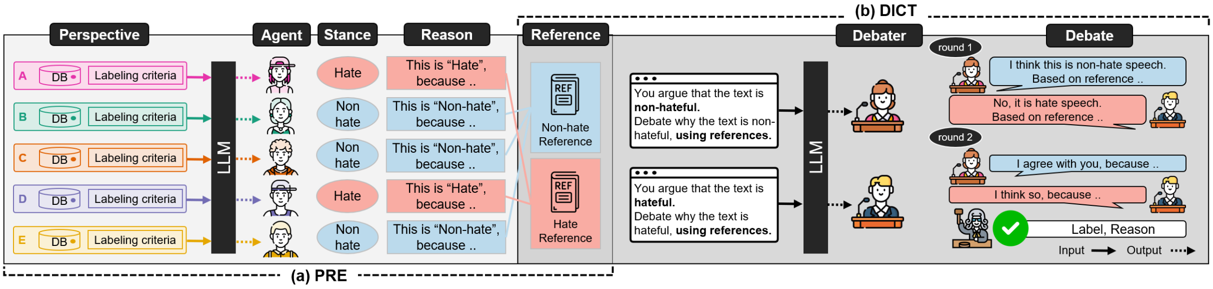
(b) DICT
Perspective
Agent
Stance
Reason
Reference
Debater
Debate
A DB Labeling criteria
B DB Labeling criteria
C DB Labeling criteria
D DB Labeling criteria
E DB Labeling criteria
LLM
Hate
Non hate
Non hate
Hate
Non hate
This is "Hate", because ..
This is "Non-hate", because ..
This is "Non-hate", because ..
This is "Hate", because ..
This is "Non-hate", because ..
REF
Non-hate Reference
REF
Hate Reference
You argue that the text is non-hateful. Debate why the text is non-hateful, using references.
You argue that the text is hateful. Debate why the text is hateful, using references.
LLM
round 1
I think this is non-hate speech. Based on reference ..
No, it is hate speech. Based on reference ..
round 2
I agree with you, because ..
I think so, because ..
Label, Reason
Input
Output
(a) PRE

Figure 2: Overview of our PREDICT framework. (a) shows agents assigned perspectives based on the unique attributes of five datasets, from which they derive their respective stances and reasons regarding the same text. These reasons are then divided into two opposing camps and stored as references. (b) shows debaters using two incongruent references to argue whether the text is hateful or not. After the debate, a judge determines the final label and reason for the text.

ased AI systems in turn reinforce social biases, creating a cycle that reinforces negative stereotypes and discriminatory attitudes towards minority groups. Recognizing this problem highlights why it is important to include agents with diverse backgrounds and biases in the PREDICT framework. By providing different perspectives, these agents can help compensate for the biases of a single agent and help improve the accuracy of hate speech detection. In further support of the PREDICT approach, Muthukrishna and Henrich (2016); Malone et al. (2009); Riedl et al. (2021) showed how groups' problem-solving abilities improve when members with diverse backgrounds and experiences participate, reinforcing the value of diverse perspectives in our framework.

3 Method

Figure 2 provides an overview of the PREDICT framework, which consists of two main phases: (a) PRE: a phase that generates agents' stances and reasons based on each refined independent perspective for a given text, and (b) DICT: a phase in which debaters simulate a debate based on their stances (i.e., hate, non-hate) and reach a consensus for hate speech detection. In the DICT phase, the agents are divided into two camps based on their stances. The debaters then combine the reasons from each camp to simulate the debate. The debate is conducted in two rounds, and finally, a judge makes the decision about hate or non-hate and generates a balanced reason that respects the arguments of each stance.

3.1 PRE: Perspective-based Reasoning

The goal of PRE (Perspective-based REasoning) is to define each agent's stance and reason to simulate a reasoning-based debate. Five agents, each assigned one of five independent perspectives, take a stance on whether the same text is hateful or not. They then generate reasons to support their respective stances. An independent perspective has two components: (1) labeling criteria and (2) similar contexts. Labeling criteria serve as the determining factor in annotating whether an unlabeled text is hateful or non-hateful. The application of these criteria results in labeled text that provides a similar context, which can be an element in establishing an independent perspective. The process of assigning an independent perspective to each agent is shown in Figure 3-(a). Based on the given perspective, each agent generates stances and reasons, as shown in Figure 3-(b).

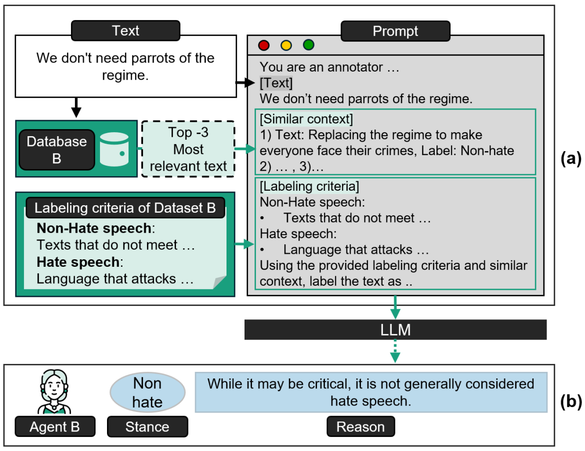
Text
We don't need parrots of the regime.
Prompt
You are an annotator …
[Text]
We don't need parrots of the regime.
[Similar context]
1) Text: Replacing the regime to make everyone face their crimes, Label: Non-hate
2) … , 3)…
Database B
Top -3 Most relevant text
Labeling criteria of Dataset B
Non-Hate speech: Texts that do not meet …
Hate speech: Language that attacks …
[Labeling criteria]
Non-Hate speech:
• Texts that do not meet …
Hate speech:
• Language that attacks …
Using the provided labeling criteria and similar context, label the text as ..
(a)
LLM
Non hate
While it may be critical, it is not generally considered hate speech.
Agent B
Stance
Reason
(b)

Figure 3: Detail of the PRE phase. (a) shows the process of extracting unique attributes from a dataset, converting them into prompts, and assigning them as perspectives to an agent. (b) shows the agent deriving a stance and a reason based on the perspective.

To reflect the labeling criteria as an independent perspective, in this paper, three researchers of this paper conducted content and thematic analy-

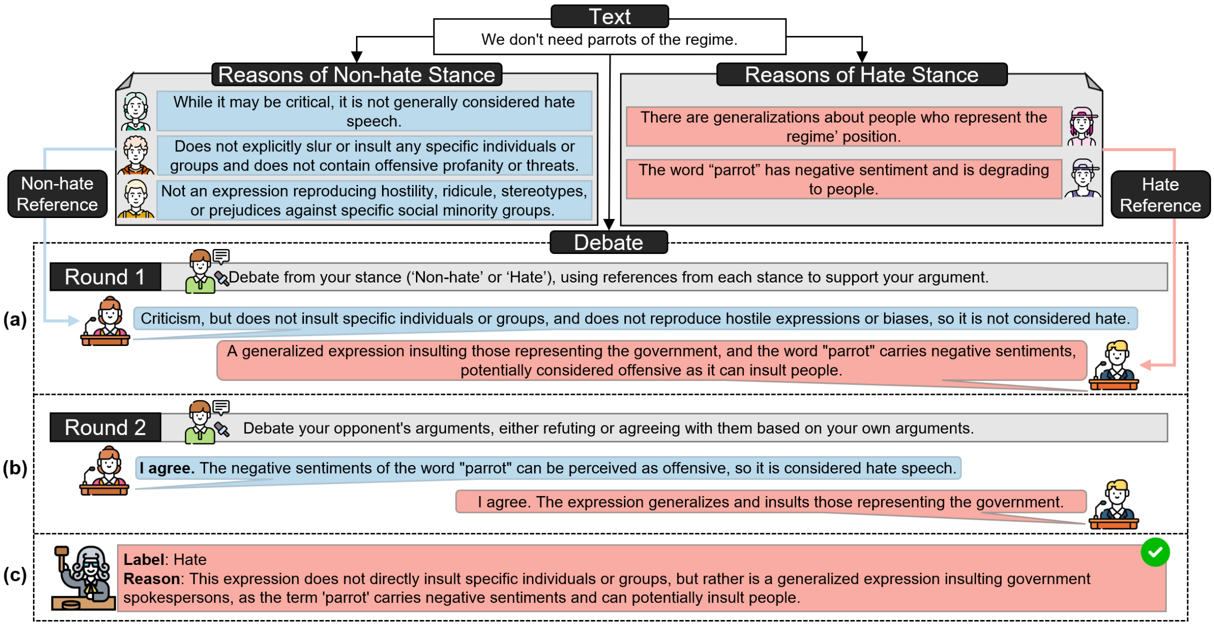
Text
We don't need parrots of the regime.
Reasons of Non-hate Stance
While it may be critical, it is not generally considered hate speech.
Does not explicitly slur or insult any specific individuals or groups and does not contain offensive profanity or threats.
Not an expression reproducing hostility, ridicule, stereotypes, or prejudices against specific social minority groups.
Reasons of Hate Stance
There are generalizations about people who represent the regime' position.
The word "parrot" has negative sentiment and is degrading to people.
Non-hate Reference
Hate Reference
Debate
(a)
Round 1
Debate from your stance ('Non-hate' or 'Hate'), using references from each stance to support your argument.
Criticism, but does not insult specific individuals or groups, and does not reproduce hostile expressions or biases, so it is not considered hate.
A generalized expression insulting those representing the government, and the word "parrot" carries negative sentiments, potentially considered offensive as it can insult people.
(b)
Round 2
Debate your opponent's arguments, either refuting or agreeing with them based on your own arguments.
I agree. The negative sentiments of the word "parrot" can be perceived as offensive, so it is considered hate speech.
I agree. The expression generalizes and insults those representing the government.
(c)
Label: Hate
Reason: This expression does not directly insult specific individuals or groups, but rather is a generalized expression insulting government spokespersons, as the term 'parrot' carries negative sentiments and can potentially insult people.

Figure 4: Detail of the DICT phase. (a) shows Debate Round 1, where debaters from both stances argue whether the given text is hateful or not, based on their respective references. (b) shows Debate Round 2, where debaters may refute or agree with their opponents' arguments, thus potentially changing their stances. (c) shows the judge determining the final label and reason.

sis on five public benchmarks in the hate speech domain based on the annotation components of each dataset (e.g., purpose of construction, labeling method) as described in Appendix A. Figure 3 shows the process of the PRE phase using Dataset B as an example. The labeling criteria in Dataset B refer to the standards for classifying texts as hateful or non-hateful. After refining the labeling criteria, the transcribed prompts were used to assign an independent perspective to each agent. To ensure stable and consistent text generation, we used the prompt frames for the persona assignment (Huang et al., 2024) as a base structure and added our independent perspectives to the prompt. The detailed prompt structure is described in Appendix D.

Then, to present the similar context as an independent perspective, the three most relevant texts from the database of a particular perspective are extracted by a cosine similarity-based search on the input text. Each extracted text consists of a sentence and a label, which serve as the basis for generating the stance and reason of each agent.

If each agent classifies as "Hate," the reason is added to the "Hate reference," and if it classifies as "Non-hate," the reason is added to the "Non-hate reference." A reference is defined as a set of arguments that can be referenced by a debater in the hate or non-hate camp to simulate a debate in the DICT phase of Section 3.2. In the PRE phase, each agent forms a unique perspective for the debate simulation based on the labeling criteria and similar contexts within the matched dataset. The pseudo-code that outlines the overall process of the PRE is described in Algorithm 1.

Algorithm 1 **PRE**: Perspective-based **RE**asoning

```
Require: Text $t$, Labeling criteria $\{A, B, C, D, E\}$, Database $\{db_A, db_B, db_C, db_D, db_E\}$
Ensure: Non-hate Reference $nh\_Ref$, Hate Reference $h\_Ref$
procedure PRE($t$, Labeling criteria, Database)
    $nh\_Ref \leftarrow []$                ▷ # Initialize non-hate reference
    $h\_Ref \leftarrow []$                 ▷ # Initialize hate reference
    for each criteria, db in zip (Labeling criteria, Database) do
        $similar\_context \leftarrow$ db.cosine_similar($t$)
        $perspective \leftarrow (criteria, similar\_context)$
        $prompt \leftarrow$ create_prompt($perspective$)
        $agent \leftarrow$ LLM($prompt$)
        # Using agent's unique perspective
        # stance (S), reason (R)
        $S, R \leftarrow agent.stance\_reason(t)$
        # Classify based on stance
        if $S =$ "Hate" then
            $h\_Ref \leftarrow h\_Ref + [R]$
        else
            $nh\_Ref \leftarrow nh\_Ref + [R]$
        end if
    end for
    return $h\_Ref, nh\_Ref$
end procedure
```

3.2 DICT: Debate using Incongruent References

The goal of DICT (Debate using InCongruenT references) is to reach a consensus on hate speech detection through multi-agent debate simulations. The overall process of the DICT phase is illustrated in Figure 4. First, before the debate rounds begin, five agents are assigned to either the hate or non-hate camp based on their stance derived in the PRE phase. Two debaters from the hate and non-hate camps, respectively, receive the references, which are sets of reasons provided by the agents assigned to their respective camps. In our framework, the debate proceeds over a total of two rounds.

At the beginning of Round 1, the moderator agent generates text to initiate the debate between the debaters of the two camps based on predetermined prompts (Figure 4-(a)). The debaters argue their stances and opinions on the input text based on their references and counter each other's arguments. The stances and arguments of both debaters are stored in the debate history.

In Round 2, the moderator agent asks each debater to refute or agree with the opponent's arguments based on predetermined prompts, as in Round 1. At this stage, each debater may revise his or her initial stance (Figure 4-(b)). Both debaters' stances and arguments at this stage are stored in the debate history. Finally, the judge agent references the debate history to reach a consensus on whether to classify the given text expression as a hate or non-hate label, and provides a balanced reason. In the DICT phase, even if the distribution of stances among the agents is skewed, the debate is conducted by the two debaters, ensuring that minority stances are represented, thus promoting fair debate simulations. The pseudo-code for the algorithm used in the DICT phase is described in Algorithm 2, and the actual prompts used in our framework are provided in Appendix E.

4 Experiments

4.1 Experimental Settings

4.1.1 Dataset

We used the Korean hate speech benchmark datasets—K-HATERS, K-MHaS, KOLD, KODORI, and UnSmile—to implement and evaluate our framework.

- K-HATERS (Park et al., 2023a): A hate speech detection corpus containing 192K news comments, each rated on a three-point Likert scale for target-specific offensiveness.
- K-MHaS (Lee et al., 2022): A multi-label hate speech dataset of 109K comments from online news, categorized into eight classes.
- KOLD (Jeong et al., 2022): A dataset of 40.4K comments from online platforms, hierarchically annotated to identify offensive language directed at individuals or groups.
- KODORI (Park et al., 2023b): A dataset of 39.5K comments from online communities and news, annotated for offensive, abusive, and sentiment.
- UnSmile (Kang et al., 2022): A multi-label hate speech dataset consisting of 35K comments from online communities and news, annotated across seven categories.

Algorithm 2 **DICT**: **D**ebate using **I**n**C**ongruen**T** references

```
Require: Text $t$, Non-Hate Reference $nh\_Ref$, Hate Reference $h\_Ref$
Ensure: Debate outcome $o$
procedure DICT($t, nh\_Ref, h\_Ref$)
    $H \leftarrow []$                                   ▷ # Initialize debate history
    $nh\_Debater \leftarrow$ InitializeDebater($t$, "Non-hate")
    $h\_Debater \leftarrow$ InitializeDebater($t$, "Hate")
    # Round 1: Argument using own reference
    # Arg: Argument
    $nh\_Arg \leftarrow nh\_Debater$.debate($nh\_Ref$)
    $H \leftarrow H.append(\{“Non-hate” : nh\_Arg\})$
    $h\_Arg \leftarrow h\_Debater$.debate($nh\_Arg, h\_Ref$)
    $H \leftarrow H.append(\{“Hate” : h\_Arg\})$
    # Round 2: Refute or Agree
    # Res: Response
    $nh\_Res \leftarrow nh\_Debater$.debate($h\_Arg, nh\_Arg$)
    $H \leftarrow H.append(\{“Non-hate” : nh\_Res\})$
    $h\_Res \leftarrow h\_Debater$.debate($nh\_Res, h\_Arg$)
    $H \leftarrow H.append(\{“Hate” : h\_Res\})$
    # Final judgment by the judge
    $J \leftarrow$ InitializeJudge($H$)
    $L, R \leftarrow J$.judgment($t$)
    $o \leftarrow \{“Label” : L, “Reason” : R\}$
    return $o$
end procedure
```

For the experiment, we randomly selected a total of 400 samples from each test dataset, consisting of 200 hate speech and 200 non-hate speech samples. Our sampling method is based on the previous research on LLM-based in-context learning (Guo et al., 2023). In this paper, these datasets are referred to as Dataset A (K-HATERS), Dataset B (K-MHaS), Dataset C (KOLD), Dataset D (KODORI), and Dataset E (UnSmile), respectively (A more detailed description of the datasets can be found in Appendix A).

4.1.2 Implementation Details

In the PRE phase, for the cosine similarity-based search, we applied OpenAI's "text-embedding-ada-002-v2" embedding to construct vector databases from the training set of each dataset, thereby retrieving similar contexts. We used FAISS (Douze et al.,

2024) as the vector database. We used the "gpt-3.5-turbo-0125" model for the five agents. To verify the consistency of stance and argument generation through assigning perspective, we performed each experiment five times and used Fleiss' Kappa as the consistency evaluation metric. In the DICT phase, we used a rule-based agent that receives prompts tailored to each role, providing predetermined text as the moderator agent, and the "gpt-3.5-turbo-0125" model for the two debater agents and the judge agent. For the robust evaluation, each experiment was repeated five times, and the performance was evaluated using both the mean and standard deviation as metrics.

4.2 Validity and Consistency of PRE

To evaluate the validity of assigning "independent perspectives" (i.e., labeling criteria, similar context) in the PRE phase, we conducted in-dataset and cross-dataset evaluations.

The in-dataset evaluation assesses how accurately agents optimized for their respective datasets detect hate speech within those datasets, compared to other agents. The cross-dataset evaluation assesses the performance of agents optimized for specific datasets in detecting hate speech in other datasets.

As shown in Table 1, each agent achieved the highest performance on its corresponding dataset (in-dataset evaluation). This indicates that the perspectives (i.e., labeling criteria, similar contexts) of the datasets optimized for each Agents_A, _B, _C, _D, and _E were effectively assigned to the agents. Note that Agent_Base was not optimized for any specific dataset and detects hate speech by zero-shot, serving as a baseline. Conversely, the performance of the agents generally decreased on datasets for which they were not optimized and was sometimes lower than the baseline (cross-dataset evaluation). For example, Agent_A showed lower performance on datasets B, C, D, and E compared to the baseline. This implies that focusing solely on the characteristics of specific datasets can reduce the level of generalization of the model.

Furthermore, we conducted an experiment to verify whether each agent consistently makes predictions based on the assigned perspective. The numbers in parentheses in Table 1 represent the Fleiss' Kappa values, which indicate the agreement of results from performing the same experiment five times. The Fleiss' Kappa interpretations provided by all agents fall into the "Almost perfect" category. The agreement among the five agents on the same text (5-Agents majority vote) shows a low level of Fleiss' Kappa, falling into the "Fair agreement" or "Moderate agreement" categories. This indicates that each agent is making consistent predictions while maintaining differences in perspectives.

Agent	**Evaluation Dataset**				
	A	**B**	**C**	**D**	**E**
Agent_Base	0.755 (0.947)	0.747 (0.976)	0.750 (0.974)	0.881 (0.978)	0.761 (0.966)
Agent_A	**0.766** **(0.939)**	0.627 (0.983)	0.624 (0.936)	0.726 (0.989)	0.612 (0.977)
Agent_B	0.657 (0.903)	**0.831** **(0.987)**	0.662 (0.936)	0.809 (0.962)	0.747 (0.962)
Agent_C	0.659 (0.879)	0.753 (0.964)	**0.808** **(0.970)**	0.860 (0.961)	0.732 (0.964)
Agent_D	0.702 (0.906)	0.758 (0.972)	0.741 (0.943)	**0.916** **(0.973)**	0.781 (0.982)
Agent_E	0.648 (0.892)	0.738 (0.961)	0.659 (0.944)	0.838 (0.951)	**0.796** **(0.971)**
5-Agents majority vote	0.743 (0.334)	0.794 (0.441)	0.767 (0.358)	0.907 (0.530)	0.808 (0.363)

Table 1: Experimental results for the analysis of in-dataset and cross-dataset performance of agents with assigned perspectives. The figures outside the brackets represent the accuracy metric, while the figures inside the brackets indicate the Fleiss' Kappa values. The "5-Agents majority vote" row indicates the majority voting accuracy and agreement for the results of five agents. Each experimental condition was repeated five times, and the average accuracy is reported. The ablation study on the two elements of perspective can be found in Appendix B.1, and interpretations of Fleiss' Kappa and details are provided in Appendix B.2

In summary, we experimentally verified the impact of perspective differences between datasets on model generalization by observing the difference between in-dataset and cross-dataset performance. Additionally, through consistency evaluation, we verified that the "independent perspectives" we defined were consistently assigned and quantitatively measured the perspective differences among the five agents. Our results demonstrate the necessity of a pluralistic approach, considering diverse labeling criteria to reach a consensus, for improving the generalization of hate speech detection models.

4.3 Effectiveness of DICT in Generalization

In this section, we evaluate the effectiveness of our proposed debate simulation, DICT, for generalized hate speech detection.

To evaluate the performance of the multi-agent debate simulation in the DICT phase, we conducted

Consensus		Evaluation Dataset				
		A	**B**	**C**	**D**	**E**
Non-Debate	**In-dataset**	0.766 (± 0.003)	0.831 (± 0.002)	0.808 (± 0.004)	0.916 (± 0.002)	0.796 (± 0.002)
	Majority voting	0.743 (± 0.002)	0.794 (± 0.003)	0.767 (± 0.001)	0.907 (± 0.002)	0.808 (± 0.002)
Debate	**Round 1**	0.745 (± 0.007)	0.794 (± 0.011)	0.753 (± 0.010)	0.910 (± 0.009)	0.795 (± 0.005)
	Rounds 1 & 2	**0.794 (± 0.003)**	**0.850 (± 0.002)**	**0.851 (± 0.004)**	**0.949 (± 0.004)**	**0.837 (± 0.009)**

Table 2: Experimental comparison of consensus methods across five agents. The "In-dataset" serves as the baseline, representing the best performing agent from each dataset in the PRE phase. This table examines whether rational consensus methods outperform the baseline. Each experimental condition was repeated five times, and the average accuracy is reported. The numbers in parentheses are the standard deviations of these trials. PREDICT with Rounds 1 & 2 debates achieved the best performance across all datasets. The ablation study for the concept of generalization can be found in Appendix B.3

comparative experiments with the non-debate methods of the in-dataset and the majority voting methods. The in-dataset and majority voting methods are the same as the one described in Table 1. The majority voting method determines the final label based on the result that receives more than half of the votes from five agents on the same text. As shown in Table 2, DICT (Debate - Rounds 1 and 2) achieved the highest performance on all public benchmarks. In contrast, despite the use of five agents, the majority voting method decreased the performance compared to the in-dataset results for all datasets except Dataset E. This result quantitatively demonstrates the limitation of not fully considering the perspective differences between datasets. We demonstrate that DICT significantly improves generalization performance by encompassing diverse perspectives and adequately reflecting minority opinions.

To evaluate the effectiveness of the setting where debaters can change their opinions, we conducted a comparative analysis of cases where only Round 1 was performed versus cases where both Rounds 1 and 2 were performed. The experimental results showed that the first case (Round 1 only) sometimes underperformed compared to the in-dataset or majority voting methods. These results indicate the limitations of Round 1, where each debater focuses solely on his or her stance without adequately considering the opinions of others.

In contrast, the second case (Rounds 1 and 2) achieved significant performance improvements. This may be due to the fact that the debate between the two debaters encourages compensation for errors (possibly due to majority bias) in Round 2 and the judge to make the right decision by revising, supplementing, or extending the arguments. Through this experiment, we quantitatively verified the superiority of our framework in generalized hate speech detection by comprehensively considering diverse opinions to reach a consensus.

Figure 4 shows a case where our framework correctly classifies hate speech by appropriately reflecting minority opinions. In Figure 4-(b), the non-hate stance debater acknowledges the overlooked aspects of his or her argument and shows respect for the opponent's stance. This allowed the judge agent to make more balanced decisions. Additionally, in Figure 4-(c), the judge comprehensively reviews the conflicting arguments from both the hate stance and non-hate stance debaters, correctly classifies the hate speech, and provides a balanced reason.

5 Discussion

5.1 Opportunities to social science research

The PREDICT framework has the potential to be applied to social science research on various topics, including hate speech, by assigning representative agents and building an environment in which the agents can interact. By modeling complex social interactions and decision-making processes, researchers could identify new patterns of behaviors and insights that may not be possible with traditional research methods in social science.

PREDICT can be extended by considering multiple agents with different characteristics, even in the same dataset. Our study considered one representative agent from one dataset and rather a simple debate condition. We can consider greater diversity in the number of agents, the number of debate phases, the level of engagement of agents, moderators, judges, and more. While it may not be necessary to find optimal parameters for the study condition, diversifying agents and environments can provide many more interesting research opportunities in social science research.

6 Conclusion

This paper proposes the PREDICT framework, a novel multi-agent-based debate simulation that employs a pluralistic approach to overcome differences in labeling criteria across datasets in the domain of hate speech detection. Experimental results on five benchmark datasets show that PREDICT performed best in cross-evaluations, demonstrating improved generalization through the integration of diverse perspectives. The results of this study indicate that the PREDICT has the potential to be applied beyond hate speech detection, providing a new direction for LLM-integrated research in the field of social sciences.

Limitations

This study improves the accuracy and generalizability of hate speech detection by structuring a multi-agent-based debate simulation. The settings of various parameters considered in the PREDICT framework have been experimentally determined, and various additional experiments are needed to achieve optimal settings. For example, the agent corresponding to each dataset can be composed of two or more agents by adding agent characteristics. In addition, there is a limit where each debater agent (i.e., non-hate, hate) can speak only once each within round 1 and round 2. Related to this, we experimentally allowed debater agents to speak and debate more than twice within each round, but we observed that this increased the number of texts entered into the prompt, causing hallucinations in the judge's judgment. This phenomenon appears to be related to the technical limitations of the current LLM rather than problems with the framework of this study. It highlights the need for continued improvement and research to overcome hallucinations due to increased prompt size.

Ethical Considerations

Our work aims to extend previous research on hate speech detection and to contribute to the resolution of social conflicts caused by hateful content. We address ethical considerations that are essential in dealing with hate speech.

One of the key concerns with using LLMs for hate speech detection is the potential for incorrect inferences due to data bias or model-generated hallucinations inherent in these models. Insufficient training data for certain groups may lead to misclassification of these groups with respect to the generation and dissemination of hate speech. Additionally, our framework may not perfectly replicate the complex human reasoning processes, and inconsistencies in LLM outputs may occur even when analyzing the same input sentence. These issues can lead to inaccurate results and potentially reinforce negative stereotypes about certain groups. We are aware of the various challenges that can arise when using LLMs in the domain of hate speech.

In response to these challenges, our research has undertaken the following efforts. We use publicly available open datasets containing hate speech. In the PRE phase, we tried to reduce the impact of the inherent bias of the LLM by guiding each agent to build a perspective based on specific labeling criteria and similar contexts. In the DICT phase, we tried to further minimize the bias effect of the LLM by ensuring that the debate simulation was based on a concrete reference. The judges in our simulations critically evaluate the arguments from both sides to ensure the accuracy of each evaluation. To ensure the reliability of our results, we repeat all experiments five times in our PREDICT framework.

Recognizing the ongoing ethical considerations required for using LLMs to detect hate speech, we aim to continue our efforts through refined prompt design, the use of various LLMs, and more robust social simulations. We believe that our continuous efforts are important to mitigate the risks associated with using LLMs for hate speech detection.

Acknowledgements

This research was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2020-II201373), and the National Research Foundation of Korea (NRF) (No.2018R1A5A7059549).

References

Teshome Mulugeta Ababu and Michael Melese Woldeyohannis. 2022. Afaan oromo hate speech detection and classification on social media. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 6612–6619.

Thomas Bose, Andreagiovanni Reina, and James AR Marshall. 2017. Collective decision-making. *Current opinion in behavioral sciences*, 16:30–34.

Tom Bourgeade, Patricia Chiril, Farah Benamara, and Véronique Moriceau. 2023. What did you learn to hate? a topic-oriented analysis of generalization in hate speech detection. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3495–3508, Dubrovnik, Croatia. Association for Computational Linguistics.

William Cai, Ro Encarnacion, Bobbie Chern, Sam Corbett-Davies, Miranda Bogen, Stevie Bergman, and Sharad Goel. 2022. Adaptive sampling strategies to construct equitable training datasets. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1467–1478.

Yuchen Cai, Ding Cao, Rongxi Guo, Yaqin Wen, Guiquan Liu, and Enhong Chen. 2024. Locating and mitigating gender bias in large language models. *arXiv preprint arXiv:2403.14409*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Curt Dudley-Marling and Mary Bridget Burns. 2014. Two perspectives on inclusion in the united states. *Global Education Review*, 1(1):14–31.

Guy Feldman. 2021. Asset-building and social inclusion: a qualitative analysis of families' perspectives. *Journal of Social Work*, 21(2):225–245.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6786–6794.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79.

Keyan Guo, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin Hu. 2023. An investigation of large language models for real-world hate speech detection. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1568–1573. IEEE.

Shi Yin Hong and Susan Gauch. 2023. Improving cross-domain hate speech generalizability with emotion knowledge. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 282–292, Hong Kong, China. Association for Computational Linguistics.

Zhengjie Huang, Pingsheng Liu, Gerard de Melo, Liang He, and Linlin Wang. 2024. Generating persona-aware empathetic responses with retrieval-augmented prompt learning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12441–12445. IEEE.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, page 126232.

Younghoon Jeong, Juhyun Oh, Jongwon Lee, Jaimeen Ahn, Jihyung Moon, Sungjoon Park, and Alice Oh. 2022. KOLD: Korean offensive language dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10818–10833, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yiping Jin, Leo Wanner, Vishakha Laxman Kadam, and Alexander Shvets. 2023. Towards weakly-supervised hate speech classification across datasets. *arXiv preprint arXiv:2305.02637*.

TaeYoung Kang, Eunrang Kwon, Junbum Lee, Youngeun Nam, Junmo Song, and JeongKyu Suh. 2022. Korean online hate speech dataset for multilabel classification: How can social science improve dataset on hate speech? *arXiv preprint arXiv:2204.03262*.

Urja Khurana, Ivar Vermeulen, Eric Nalisnick, Marloes Van Noorloos, and Antske Fokkens. 2022. Hate speech criteria: A modular approach to task-specific hate speech definitions. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 176–191.

Abhishek Kumar, Sarfaroz Yunusov, and Ali Emami. 2024. Subtle biases need subtler measures: Dual metrics for evaluating representative and affinity bias in large language models. *arXiv preprint arXiv:2405.14555*.

Jean Lee, Taejun Lim, Heejun Lee, Bogeun Jo, Yangsok Kim, Heegeun Yoon, and Soyeon Caren Han. 2022. K-MHaS: A multi-label hate speech detection dataset in Korean online news comment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3530–3538, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Florian Ludwig, Klara Dolos, Torsten Zesch, and Eleanor Hobley. 2022. Improving generalization of hate speech detection systems to novel target groups via domain adaptation. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 29–39.

Thomas W Malone, Robert Laubacher, and Chrysanthos Dellarocas. 2009. Harnessing crowds: Mapping the genome of collective intelligence.

Albert E Mannes, Jack B Soll, and Richard P Larrick. 2014. The wisdom of select crowds. *Journal of personality and social psychology*, 107(2):276.

Tian Xiang Moy, Mafas Raheem, and Rajasvaran Logeswaran. 2021. Hate speech detection in english and non-english languages: A review of techniques and challenges. *Technology*.

Michael Muthukrishna and Joseph Henrich. 2016. Innovation in the collective brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1690):20150192.

Ayushi Nirmal, Amrita Bhattacharjee, Paras Sheth, and Huan Liu. 2024. Towards interpretable hate speech detection using large language model-extracted rationales. *arXiv preprint arXiv:2403.12403*.

Chaewon Park, Soohwan Kim, Kyubyong Park, and Kunwoo Park. 2023a. K-HATERS: A hate speech detection corpus in Korean with target-specific ratings. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14264–14278, Singapore. Association for Computational Linguistics.

San-Hee Park, Kang-Min Kim, O-Joun Lee, Youjin Kang, Jaewon Lee, Su-Min Lee, and SangKeun Lee. 2023b. "why do I feel offended?" - Korean dataset for offensive language identification. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1142–1153, Dubrovnik, Croatia. Association for Computational Linguistics.

Sagi Pendzel, Tomer Wullach, Amir Adler, and Einat Minkov. 2024. Generative ai for hate speech detection: Evaluation and findings. *Regulating Hate Speech Created by Generative AI*, page 54.

Ramya Ramalingam, Nicolas Espinosa Dice, Megan L Kaye, and George D Montañez. 2022. Bounding generalization error through bias and capacity. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Christoph Riedl, Young Ji Kim, Pranav Gupta, Thomas W Malone, and Anita Williams Woolley. 2021. Quantifying collective intelligence in human groups. *Proceedings of the National Academy of Sciences*, 118(21):e2005737118.

Georgios Rizos, Konstantin Hemker, and Björn Schuller. 2019. Augment to prevent: short-text data augmentation in deep learning for hate-speech classification. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 991–1000.

Pratik S Sachdeva, Renata Barreto, Claudia von Vacano, and Chris J Kennedy. 2022. Assessing annotator identity sensitivity via item response theory: A case study in a hate speech corpus. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1585–1603.

Vijay Sarthy Mysore Sreedhara and Gregory Mocko. 2015. Control of thermoforming process parameters to increase quality of surfaces using pin-based tooling. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 57113, page V004T05A016. American Society of Mechanical Engineers.

Vighnesh Subramaniam, Antonio Torralba, and Shuang Li. 2023. Debategpt: Fine-tuning large language models with multi-agent debate supervision.

Alice Tontodimamma, Eugenia Nissi, Annalina Sarra, and Lara Fontanella. 2021. Thirty years of research into hate speech: topics of interest and their evolution. *Scientometrics*, 126:157–179.

Zeerak Waseem, James Thorne, and Joachim Bingel. 2018. Bridging the gaps: Multi task learning for domain transfer of hate speech detection. *Online harassment*, pages 29–55.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.

Tomer Wullach, Amir Adler, and Einat Minkov. 2021. Fight fire with fire: Fine-tuning hate detectors using large samples of generated hate speech. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4699–4705, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. 2023. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7572–7590, Singapore. Association for Computational Linguistics.

Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. 2023. HARE: Explainable hate speech detection with step-by-step reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5490–5505, Singapore. Association for Computational Linguistics.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.

Xianbing Zhou, Yang Yong, Xiaochao Fan, Ge Ren, Yunfeng Song, Yufeng Diao, Liang Yang, and Hongfei Lin. 2021. Hate speech detection based on sentiment knowledge sharing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7158–7166.

A Dataset Taxonomy Table

	K-HATERS	K-MHaS	KOLD	KODOLI	UnSmile
Purpose	To evaluate the degree of offensiveness with target-specific ratings using a three-point Likert scale.	To detect hate speech considering Korea's social and historical context to effectively handle Korean language patterns.	Employs a hierarchical taxonomy to improve model accuracy by identifying the type, target, and span of offensive language in Korean, reflecting cultural and linguistic nuances.	To enhance the detection of offensive Korean language by integrating abusive language detection and sentiment analysis.	To enhance the detection of hate speech within the context of Korea's diverse cultural backgrounds, using a multi-label approach based on social science.
Labeling Process	(A) Label fine-grained ratings (Insult, Swear word, Obscenity, Threat) (B) Label target-specific ratings (Gender, Age, Race/Origin, Religion, Politics, Job, Disability, Individual, Others) (C) Offensive and Target Rationale	(A) Binary classification (Hate Speech, Not Hate Speech) (B) 8 fine-grained hate speech (Politics, Origin, Physical, Age, Gender, Religion, Race, Profanity)	(A) Offensive Language Detection (OFF, NOT) (B) Target Type Categorization of Offensive Language (UNT, IND, GRP, OTH) (C) Target Group Identification of Group Targeted offensive Language (Target group attribute, Target Group)	(A) Offensive Language Identification (OFFEN, LIKELY, NOT) (B) Abusive Language Detection (ABS, NON) (C) Sentiment Analysis (POS, NEG, NEU)	(A) Multi-label (Race/Nationality, Religion, Regionalism, Ageism, Women/Family, Sexual Minorities, Male, Profanity, Extra hate speech, Clean)
Label class	- Not Offensive (normal) - Offensive (i.e., offensive, implicit hate speech, explicit hate speech)	- Not Hate Speech - Hate Speech	- Not Offensive - Offensive (i.e., offensive, hate speech)	- Not Offensive (normal) - Offensive (i.e., likely offensive, offensive)	- Not Hate Speech - Hate Speech

	K-HATERS	K-MHaS	KOLD	KODOLI	UnSmile
Feature	1. Only offensive remarks towards target groups (e.g., (Gender, Age, Race/Origin, Religion, Politics, Job, Disability) are considered hate speech (in our method, hate speech is also classified as offensive). 2. Hate speech is further categorized based on the presence or absence of expression (as determined by rationale).	1. Top 5 keywords associated with each fine-grained label, which were provided to annotators during labeling as a reference.	1. Only offensive remarks towards target groups (e.g., Gender, Race, Political Affiliation, Religion, Miscellaneous) are considered hate speech (in our method, hate speech is also classified as offensive). 2. Untargeted offensive remarks are classified as offensive.	1. Focuses on detecting overall offensiveness and hate speech without providing specific targeting criteria or details for the target group. 2. Enables more accurate assessment of offensiveness by identifying the intended emotional state of comments through sentiment analysis. 3. Evaluates offensiveness by analyzing both intention and abusive language	1. Self-deprecation is not considered as hate speech. 2. Derogatory and discriminatory remarks that refer to the group to which the multi-label class belongs. 3. Stereotypes about the target in the multi-label class. 4. Remarks that fixate the target's characteristics or tendencies in the multi-label class to a specific stereotype.

Table 3: Five benchmark datasets with distinct characteristics. Three researchers conducted content and thematic analysis on these public hate speech benchmarks, focusing on the annotation components of each dataset.

B Ablation study

B.1 Ablation study for components of perspective in PRE

We conducted an ablation study to evaluate the impact of including two key components, labeling criteria and similar context, when assigning independent perspectives. The results show that applying only one of these components does not consistently lead to the highest in-dataset performance for the agent. In contrast, when both "Labeling criteria" and "Similar context" were applied, all agents achieved the highest performance across all in-dataset evaluations. These results indicate that both labeling criteria and similar context play a key role in perspective assignment, which is crucial for accurately reflecting the characteristics of each dataset. To ensure robustness, we conducted each experiment five times, and the averages are reported.

Base dataset	Perspective		Evaluation Dataset				
			Accuracy				
	Labeling criteria	Similar context	A	B	C	D	E
A	✗	✓	0.614	0.647	0.638	0.733	0.597
	✓	✗	0.747	0.805	0.708	0.892	0.775
	✓	✓	**0.766**	0.627	0.624	0.726	0.612
B	✗	✓	0.680	0.808	0.702	0.827	0.752
	✓	✗	0.699	0.819	0.685	0.868	0.767
	✓	✓	0.657	**0.831**	0.662	0.809	0.747
C	✗	✓	0.720	0.750	0.785	0.857	0.801
	✓	✗	0.731	0.791	0.761	0.895	0.784
	✓	✓	0.659	0.753	**0.808**	0.860	0.732
D	✗	✓	0.670	0.775	0.720	0.844	0.793
	✓	✗	0.726	0.805	0.717	0.906	0.793
	✓	✓	0.702	0.758	0.741	**0.916**	0.781
E	✗	✓	0.667	0.755	0.682	0.862	0.775
	✓	✗	0.683	0.797	0.697	0.865	0.793
	✓	✓	0.648	0.738	0.659	0.838	**0.796**

Table 4: To investigate whether including two key components of independent perspective allows an agent to accurately reflect the unique characteristics of each dataset. The metric evaluated is the agent's accuracy.

B.2 Consistency evaluation using Fleiss' Kappa in PRE

In this study, each Agents_A,_B,_C,_D, and _E conducted five repeated experiments on the same text across datasets (A, B, C, D, E). To assess the consistency of the results from each agent's five repeated trials, we employed the Fleiss' Kappa statistic. While Fleiss' Kappa is generally used to assess the agreement among three or more raters (Ababu and Woldeyohannis, 2022), in this context, it was adapted to evaluate the consistency of results across the independently repeated experiments conducted by each agent. Fleiss kappa values range from 0 to 1, with higher values indicating greater consistency in the agent's results. Table 5 shows the detailed interpretation of Fleiss' kappa adopted Sreedhara and Mocko (2015). The Fleiss Kappa statistic is calculated using the following equation:

$$\kappa = \frac{P - P_e}{1 - P_e} \tag{1}$$

where P is the proportion of agreement by chance, and P_e is the proportion of agreement by analytical reasoning.

Fleiss' Kappa	Interpretation
<0.00	Poor agreement
0.00 to 0.20	Slight agreement
0.21 to 0.40	Fair agreement
0.41 to 0.60	Moderate agreement
0.61 to 0.80	Substantial agreement
0.81 to 1.00	Almost perfect

Table 5: Interpretation of Fleiss kappa values

B.3 Ablation study for Generalization in DICT

Debate (Rounds 1&2)	Evaluation Dataset				
	Accuracy (Standard Deviations)				
	A	B	C	D	E
4-Agents	0.787 (±0.002)	0.851 (±0.002)	0.842 (±0.003)	0.935 (±0.003)	0.822 (±0.004)
5-Agents	0.794 (±0.003)	0.850 (±0.002)	0.851 (±0.004)	0.949 (±0.004)	0.837 (±0.009)

Table 6: Comparative results of "Debate - Rounds 1 and 2" showing the impact on performance when agents optimized for each respective evaluation dataset are excluded. The table presents the accuracy and standard deviations for both 4-Agents and 5-Agents setups across five evaluation datasets (A, B, C, D, E). This method demonstrates our framework's robustness and generalization capabilities by testing it without agents specifically optimized for the evaluation datasets.

In our study, the "Generalizability" we intended was that in the evaluation of a particular dataset, the multi-agent-based debate simulations would allow us to incorporate diverse perspectives to mitigate the overfitting of a single agent optimized for that dataset. The concept of "Generalizability" is used to describe the capacity to maintain performance on datasets that were not used during training. This indicates that the model does not solely rely on specific training data but operates effectively across various datasets.

To test the generalizability of our framework rigorously, we conducted an ablation study to evaluate the impact of agents optimized for specific datasets on the overall system performance. In this study, a specific dataset was selected as the test data, and the agent optimized in that dataset was excluded and then its performance was evaluated. For instance, when using the Dataset A as the test data, the agent optimized in Dataset A was removed and experiments were conducted under these conditions.

Table 6 presents the results of this experiment, which show similar performance to those of the "Debate - Rounds 1 and 2" experiment using the five agents shown in Table 2. The results confirm that removing dataset-optimized agents does not significantly reduce the ability of our framework to detect hate speech. Consequently, these findings provide rigorous support for the generalizability of our framework. To ensure robustness, we conducted each experiment five times, and the averages are reported.

C Examples of PREDICT

The following are five examples of the PREDICT framework on hate speech detection.

Text	Maybe we should not say anything. Let's go to aphasia society.
Ground truth	Non-hate
Agent A	Stance: Hate Reason: The text is categorized as implicitly hateful towards people with disabilities because it uses the term "aphasia".
Agent B	Stance: Non-hate Reason: The text does not attack or diminish any race, religion, gender, etc., and is considered "Non-hate Speech" based on the given criteria.
Agent C	Stance: Non-hate Reason: The text is considered Non-Hateful because there is no explicit or implicit offensive content such as hate speech, profanity, insults, or threats.
Agent D	Stance: Hate Reason: The text contains the phrase "Let's go to aphasia society" which could be interpreted as a rude joke using sarcasm.
Agent E	Stance: Non-hate Reason: The text does not reproduce stereotypes or prejudices, and no offensive or discriminatory statements identify a group.
PREDICT	**Non-hate side** The text is non-hateful because it does not attack or diminish a race, religion, gender, etc., and there is no explicit or implicit offensive content such as hate speech, profanity, insults, or threats. **Hate side** It is classified as implicitly hate speech against a disability (aphasia), which can be interpreted as a rude joke using sarcasm. **Non-hate side** I disagree, claiming that it is implicitly hateful disability. Still, nothing is explicitly offensive or hateful about it, and it could be interpreted as a simple joke using satire. **Hate side** I disagree, the text can be interpreted as implicitly disability hateful, and it cannot be interpreted as a simple humorous joke that is non-hateful. **Judge** Label: Non-hate, Reason: The text is judged to be non-hateful because it does not attack or diminish any race, religion, gender, etc., and has no explicit or implicit hateful content, but it does have the potential to be implicitly hateful of a disability (aphasia), so caution is required.

Text	One idiot and the media are falling for it.
Ground Truth	Hate
Agent A	Stance: Hate Reason: It uses the slur “idiot” and contains derogatory expressions.
Agent B	Stance: Non-hate Reason: Non-hateful because the text does not attack or diminish a specific individual or group.
Agent C	Stance: Hate Reason: Determined to be hateful because it contains profanity and ridicule.
Agent D	Stance: Hate Reason: It uses derogatory terms to degrade the media and is classified as offensive because it contains profanity.
Agent E	Stance: Non-hate Reason: Does not contain expressions that reproduce hostility, ridicule, sarcasm, stereotypes, or prejudice against minority groups.
PREDICT	**Non-hate side** Non-hateful because the text does not attack or diminish a specific individual or group, and does not contain language that reproduces hostility, ridicule, satire, stereotypes, or prejudice against minority groups. **Hate side** It is hateful because it uses the profanity “idiot,” uses degrading language about the media, and contains ridicule. **Non-hate side** I disagree, the text is non-hateful because it does not use profanity or demeaning language, and it does not attack or diminish a specific individual or group. **Hate side** The text contains the swear word “idiot” and uses degrading language that demeans the media; therefore, it is classified as hate speech. **Judge** Label: Hate, Reason: Although not directed at a specific individual or group, the text’s use of the word “idiot” and degrading language about the media could be interpreted as hate speech.

Text	After all, men should be young.... Young & Rich & Handsome Peace!
Ground Truth	Hate
Agent A	Stance: Hate Reason: Implies implicit hate speech and stereotypes against men.
Agent B	Stance: Non-hate Reason: The text expresses a taste for men and does not attack or diminish a specific individual or group.
Agent C	Stance: Non-hate Reason: Does not explicitly contain derogatory or hate speech.
Agent D	Stance: Hate Reason: Contains sexist comments using gender stereotypes against men.
Agent E	Stance: Non-hate Reason: Contains no offensive or discriminatory comments about a specific social minority group and is a general discussion about men.
PREDICT	**Non-hate side** The text is non-hateful because it is a statement of taste about men. The text does not attack a specific individual or group, and it does not contain derogatory or hateful expressions. **Hate side** This is hate speech. The speech is deemed hateful due to its use of sexist remarks about men and gender stereotypes. **Non-hate side** I disagree. The statement does not contain any sexist remarks against men. It is simply an expression of preference. **Hate side** I disagree. The statement uses gender stereotypes to make sexist remarks against men, so it is not merely an expression of preference. **Judge** Label: Hate, Reason: This statement is not just an expression of preference. It uses gender stereotypes to make sexist remarks against men.

Text	Is this even a country? Jaeang[1]?
Ground Truth	Hate
Agent A	Stance: Non-hate Reason: No offensive language. No insults, profanity, obscenity, or threats.
Agent B	Stance: Hate Reason: The content demeans and attacks a specific person (Moon Jae-in).
Agent C	Stance: Non-hate Reason: Does not contain insults or profanity and is simply a question of curiosity.
Agent D	Stance: Non-hate Reason: No profanity or hate speech, no direct insults to minority groups.
Agent E	Stance: Hate Reason: Contains demeaning remarks about a specific person.
PREDICT	**Non-hate side** The text is non-hateful because it contains no offensive or abusive language and is simply a form of questioning about a social situation. **Hate side** The text is hateful because it contains remarks that demean and attack a specific person (Moon Jae-in). **Non-hate side** I agree. The text is judged to be hateful because it contains statements that demean and attack a specific person (Moon Jae-in). **Hate side** I agree with my opponent, the text demeans a specific person (Moon Jae-in) by calling him a “disaster”. **Judge** Label: Hate, Reason: I find the phrase “Jaeang?” to be hateful because it contains language that demeans and attacks a specific person (Moon Jae-in).

[1] In Korea, the word for “disaster” is pronounced “jaeang”, which phonetically resembles the name of former President Moon Jae-in. This similarity in sound has been used to refer to his presidency in a derogatory manner.

Text	Shouldn’t Song Gain be on “Mom’s Diary: My Ugly Duckling[2]”?
Ground Truth	Non-hate
Agent A	Stance: Non-hate Reason: Contains no offensive or hateful language, and no language likely to offend readers.
Agent B	Stance: Non-hate Reason: Non-hate speech that explicitly attacks or degrades a specific individual or group based on attributes such as physical appearance, religion, gender, etc.
Agent C	Stance: Hate Reason: Demeans or ridicules a specific individual and is offensive.
Agent D	Stance: Non-hate Reason: Non-hateful because it does not contain explicit insults, threats, or offensive language.
Agent E	Stance: Hate Reason: Contains derogatory language that refers to a specific individual (Song Gain) as an “ugly duckling” and is mocking.
PREDICT	**Non-hate side** The sentence does not contain any offensive or hateful language, nor does it explicitly degrade a specific group of individuals or contain language that is likely to offend readers. **Hate side** The text is hateful because it is derogatory and ridicules a specific individual, Song Gain, by calling him an “ugly duckling”. **Non-hate side** I disagree, it is not clear that the intent is to demean or insult anyone, and it can be interpreted as merely questioning the need for Song Gain to be on a particular program. **Hate side** I agree with the other side. The remark is not intended to attack or insult a specific person. **Judge** Label: Non-hate, Reason: The comment is not intended to attack or insult any specific person and simply raises questions about the need for the Song Gain to be on a particular program.

[2] “Mom’s Diary: My Ugly Duckling” is a television variety show where the daily lives of unmarried celebrities are observed and chatted about by their mothers in a studio setting.

D PRE prompts

Prompts reflecting perspectives from each dataset. Based on the differences between datasets described in Appendix A, this study extracted characteristics from five benchmark hate speech datasets and created prompts based on these characteristics. The outputs were initially aligned with each dataset's labeling, but they were ultimately standardized to "Non-hate" or "Hate" categories in our study.

K-HATERS prompt

You are an annotator trained on the labeling criteria of labels. Use the following pieces of retrieved similar context to annotate the given text. If you don't know the answer, just say that you don't know. Keep the answer concise.

Text: '{text}'

1) Labeling Criteria

Not Offensive:

- Does not contain any offensive expressions including both fine-grained and target-specific ratings.
- No indication of offense toward any target or individual.

Offensive:

- Contains explicitly offensive expressions that are toxic and likely to annoy readers.
- Covers implicit hate expressions such as sarcasm and stereotypes.
- Fine-Grained Ratings: Insult, Swear word, Obscenity, Threat.
- Target-Specific Ratings: Gender, Age, Race/Origin, Religion, Politics, Job, Disability, Individual, Others.
- Hate Speech: Classified as hate speech only if offensive towards targets.
- Hate speech is further categorized based on the presence or absence of expression.

2) Use the Similar context for reference only

Similar context: {context}

Using the provided labeling criteria and similar context, label the text as either "Not Offensive" or "Offensive". Explain your labeling decision in one sentence, aligning strictly with the provided criteria.

Now please output your answer in JSON format, with the format as follows: {{"Label": "Not Offensive or Offensive", "Reason": " "}}

K-MHaS prompt

You are an annotator trained on the labeling criteria of labels. Use the following pieces of retrieved similar context to annotate the given text. If you don't know the answer, just say that you don't know. Keep the answer concise.

Text: '{text}'

1) Labeling Criteria

Not Hate Speech:

- Texts that do not meet the criteria for hate speech.
- Free of any form of profanity or offensive language.
- Does not target individuals or groups based on specific target attributes.

Hate Speech:

- Language that attacks or diminishes individuals or groups based on specific target attributes.
- Target Attributes: Origin, Physical, Politics, Age, Gender, Religion, Race.
- Includes simple profanity.

2) Use the Similar context for reference only

Similar context: {context}

Using the provided labeling criteria and similar context, label the text as either "Not Hate Speech" or "Hate Speech". Explain your labeling decision in one sentence, aligning strictly with the provided criteria.

Now please output your answer in JSON format, with the format as follows: {{"Label": "Not Hate Speech or Hate Speech", "Reason": " "}}

KOLD prompt

You are an annotator trained on the labeling criteria of labels. Use the following pieces of retrieved similar context to annotate the given text. If you don't know the answer, just say that you don't know. Keep the answer concise.

Text: '{text}'

1) Labeling Criteria

Not Offensive:

- Texts that do not meet the offensive criteria.
- Free of untargeted profanity and targeted offenses like insults and threats, which can be implicit or explicit.

Offensive:

- Contains untargeted profanity (offensive remarks) or targeted offenses such as insults and threats.
- Target Type: UNT (Untargeted), IND (Individual), GRP (Group), OTH (Others).
- Target Group Attribute: Gender & Sexual Orientation, Race, Ethnicity & Nationality, Political Affiliation, Religion, Miscellaneous.
- Hate Speech: Classified as hate speech only if offensive towards the Target Group.

2) Use the Similar context for reference only

Similar context: {context}

Using the provided labeling criteria and similar context, label the text as either "Not Offensive" or "Offensive". Explain your labeling decision in one sentence, aligning strictly with the provided criteria.

Now please output your answer in JSON format, with the format as follows: {{"Label": "Not Offensive or Offensive", "Reason": " "}}

KODOLI prompt

You are an annotator trained on the labeling criteria of labels. Use the following pieces of retrieved similar context to annotate the given text. If you don't know the answer, just say that you don't know. Keep the answer concise.

Text: '{text}'

1) Labeling Criteria

Not Offensive:

- Comments that do not contain direct or indirect offense.
- Free of profanity, targeted offense, and abusive language even if unintentional.

Offensive:

- Contains non-acceptable language or a targeted offense (group or individual).
- Can include insults, threats, and sexual harassment.
- May hide offensive intentions behind sarcasm, irony, or backhanded jokes.
- Sentiment Analysis: Used to understand the emotional state intended by the comment.
- Evaluation of Offensiveness: Analyzes both intention and abusive language.

2) Use the Similar context for reference only

Similar context: {context}

Using the provided labeling criteria and similar context, label the text as either "Not Offensive" or "Offensive". Explain your labeling decision in one sentence, aligning strictly with the provided criteria.

Now please output your answer in JSON format, with the format as follows: {{"Label": "Not Offensive or Offensive", "Reason": " "}}

UnSmile prompt

You are an annotator trained on the labeling criteria of labels. Use the following pieces of retrieved similar context to annotate the given text. If you don't know the answer, just say that you don't know. Keep the answer concise.

Text: '{text}'

1) Labeling Criteria

Not Hate Speech:

- Texts that do not meet the criteria for hate speech.
- Includes self-deprecating expressions.

Hate Speech:

- Reproduces hostility, ridicule, or prejudice against specific social minority groups (Race/Nationality, Religion, Regionalism, Ageism, Women/Family, Sexual Minorities, Male, Profanity, Extra hate speech).
- Explicit Elements: Derogatory or discriminatory statements that clearly identify the social minority groups.
- Includes simple profanity.
- Stereotypes and statements that fixate on conventional beliefs.
- Note: Self-deprecation is not considered hate speech.

2) Use the Similar context for reference only

Similar context: {context}

Using the provided labeling criteria and similar context, label the text as either "Not Hate Speech" or "Hate Speech". Explain your labeling decision in one sentence, aligning strictly with the provided criteria.

Now please output your answer in JSON format, with the format as follows: {{"Label": "Not Hate Speech or Hate Speech", "Reason": " "}}

E DICT prompt

Debate prompts for the DICT phase, associated with each line of Algorithm 2.

Initialize Prompt: Default Prompt (Algorithm 2: 3-4 line)

“NonHate_meta_prompt” *System Prompt*: “You are a debater. Welcome to a debate about determining whether a text is hateful. The topic of debate is: Is the text ‘##text##’ hateful? You argue that the text is non-hateful.”

“Hate_meta_prompt” *System Prompt*: “You are a debater. Welcome to a debate about determining whether a text is hateful. The topic of debate is: Is the text ‘##text##’ hateful? You argue that the text is hateful.”

Round 1: Argument Prompt (Algorithm 2: 7-9 line)

“NonHate_prompt_1” *User Prompt*: “You think the text is non-hateful. Reference: ‘##Non_Hate_Reference##’. Based on the Reference, give a reason why you think the sentence is non-hateful. If you don’t have a reference, give a reason why you think it is non-hateful. Please do so in one sentence.”

“Hate_prompt_1” *User Prompt*: “The non-hate side’s argument is: ##non_arg##. You disagree with the non-hate side’s argument. You think the text is hateful. Reference: ‘##Hate_Reference##’. Based on the Reference, give a reason why you think the sentence is hateful. If you don’t have a reference, give a reason why you think it is hateful. Please do so in one sentence.”

Round 2: Rebuttal or Agreement Prompt (Algorithm 2: 13-15 line)

“NonHate_arg_prompt” *Assistant Prompt*: “Your argument is: ##non_arg##.”

“NonHate_prompt_2” *User Prompt*: “The hate side’s argument is: ##hate_arg##. Based on your argument, agree or rebut the hate side’s argument and explain your reason in one sentence.”

“Hate_arg_prompt” *Assistant Prompt*: “Your argument is: ##hate_arg##.”

“Hate_prompt_2” *User Prompt*: “The non-hate side’s argument is: ##non_res##. Based on your argument, agree or rebut the non-hate side’s argument and explain your reason in one sentence.”

Final Judgement Prompt (Algorithm 2: 18-19 line)

“Judge_prompt_1” *Assistant Prompt*: “The debate on whether the text is hateful has concluded. The debate covered is: ##history##. As a judge, assess the debaters’ arguments based on the following criteria: Accuracy: How well the argument captures the non-hatefulness or hatefulness of the text.”

“Judge_prompt_2” *User Prompt*: “Consider both sides fairly to maintain a balanced perspective and make a broad judgment. Give your final judgment on whether the following text is non-hateful or hateful: ‘##text##’. Summarize your reasons in one sentence and output your decision in the following JSON format: {\“Label\”: \“Non-hate or hate\”, \“Reason\”: \“\”}. Ensure to output strictly in JSON format; include only the relevant content.”